The International Journal of Engineering and Science (IJES)
|| Volume || 10 || Issue || 5 || Series II || Pages || PP 53-59|| 2021 ||
ISSN (e): 2319-1813 ISSN (p): 20-24-1805

Highly Accurate Face Location Based on YOLO for Ultra High Resolution Video Based on HEVC

Wei-Chen Li, Ke-Nung Huang, Chih-Cheng Huang, Chou-Chen Wang
Department of Electronic Engineering, I-Shou University, Kaohsiung, Taiwan
Corresponding Author: Chou-Chen Wang

--**ABSTRACT**--

High efficiency video coding (HEVC) is a very popular coding standard since it can support up to 4k or 8k high resolution video. Face location is one of the most important tasks for intelligent video surveillance (IVS) system based on HEVC. It is a fact that an accurate face location facilitates subsequent face recognition and security system. Recently, the YOLO (you only look ones) is a very effective object location method, but it works with low resolution images. Since the modified version of YOLO (YOLOv2) location system limits the resolution of input image up to square ratio of 608×608 pixels, the input image needs to be resized to a fixed size. This will lead to a large reduction in the accuracy of face location for HEVC videos. In order to improve accuracy of face location in higher resolution images, we propose an overlapping crop method to obtain two square ratio image to match YOLOv2 limitation of image resolution. Therefore, we can achieve higher accuracy of face location in pixel domain for HEVC videos. When the proposed overlapping crop method is applied in 4K ultra high definition (UHD) images, we can reach $Recall_{500} = 95\%$ *and* $Recall_{1000} = 95\%$ *by QP=32. However, the recall rate which directly inputs 4K image to YOLOv2 only can reach* $Recall_{500} = 80\%$ *and* $Recall_{1000} = 83\%$*, respectively. Therefore, the proposed method can finish a higher accuracy of face location in high resolution videos.*

KEYWORDS*–Face location, HEVC, YOLO, convolutional neural network.*

Date of Submission: 12-05-2021 Date of Acceptance: 25-05-2021

I. INTRODUCTION

Nowadays, high efficiency video coding (HEVC) is the most commonly used video formats for recording, compression and distribution of videos. This is because the demand for high resolution video or ultra high definition (UHD) video has rapidly increased in a number of industries, especially in intelligent video surveillance, video conference and live streaming [1-2]. On the other hand, intelligent video analytics (IVA) including moving object detection, segmentation, classification and recognition are the most important tasks for intelligent video surveillance (IVS) system [3]. Face location is one of the most important tasks for IVS system based on HEVC. It is a fact that an accurate face location facilitates subsequent face recognition and security system. Recently, the series of YOLO (you only look ones) [4-5] are very effective object location methods based on ImageNet [6] and PASCAL VOC 2007 [7] datasets, but they generally work with low resolution images. For a high resolution image, it will lose a lot of detail when resized into a low resolution image. Therefore, we will lose many detections of human figure in lower resolution of the models as compared to the original resolution due to resizing.

The success of YOLO networks [4-5] inspired many recent works, targeting real-time performance for face detection and location [8-9]. However, since object detection system of YOLOv2 limits the resolution of input image up to square ratio of 608×608 pixels, the input UHD image needs to be resized to a fixed size. This will lead to a large reduction in the accuracy of face location for HEVC videos. In addition, since most of existing computer vision methods used in commercial IVS systems operate in a pixel domain, this means that IVS need to analyze video after fully decoding HEVC bitstreams or compressed files. Because the new generation of IVS system will employ UHD images, resulting in a high computational load for IVA in the pixel domain. Therefore, it is difficult to achieve a real-time and accurate face location for IVS system. In order to improve accuracy of face location in UHD images, we propose an overlapping crop method to obtain two square ratio image to fetch through the limitation of YOLOv2.

The remainder of this paper is organized as follows. In Section II we take a brief reviews of HEVC decoder and YOLO series. Section III elaborates the proposed overlapping crop method. The experimental results are presented in Section IV. Finally, Section V summarizes our conclusions.

II. BRIEF REVIEWS OF HEVC AND YOLO

2.1 HEVC decoder

Since the rapid development of electronic technology, the UHD resolution of 4K×2K (or 8K×4K) have been became the main video applications today. Today, HEVC is a very popular video codec standard for 4K/8K videos. This is because HEVC standard can provide better video quality with a lower bitrate.

HEVC decoding procedure is shown in Fig.1. HEVC decoder mainly consists of many modules including network adaptation layer (NAL) header decoding, type decoding, entropy decoding (ED), inverse quantization and inverse integer cosine transform (IQ/IT), intra frame prediction (IFP), motion compensation (MC), sample adaptive offset (SAO)and de-blocking filter (DF) which integrated loop picture filter (IPF), and other modules. The high-level decoding syntax architecture of HEVC is similar to that of H.264 [10-11]. The two layer structures of NAL and video coded layer (VCL) have been reserved. The sequence parameter set (SPS) and picture parameter set (PPS) structures have been completed with a new video parameter set (VPS) structure. On the other hand, HEVC bitstream is an ordered sequence of the syntax elements. Each syntax element is placed into a logical packet called a NAL unit (NALU).VPS, SPS and PPS contain quantization parameter (QP) and general video parameters. They provide a robust mechanism for conveying data that are essential to the decoding process. They can be either a part of bitstream or can be stored separately. Therefore, NAL header decoding is simple processing because it only take a look up table and get the corresponding parameters. As a result, we can perform decoding modules to obtain the high quality decoded images.

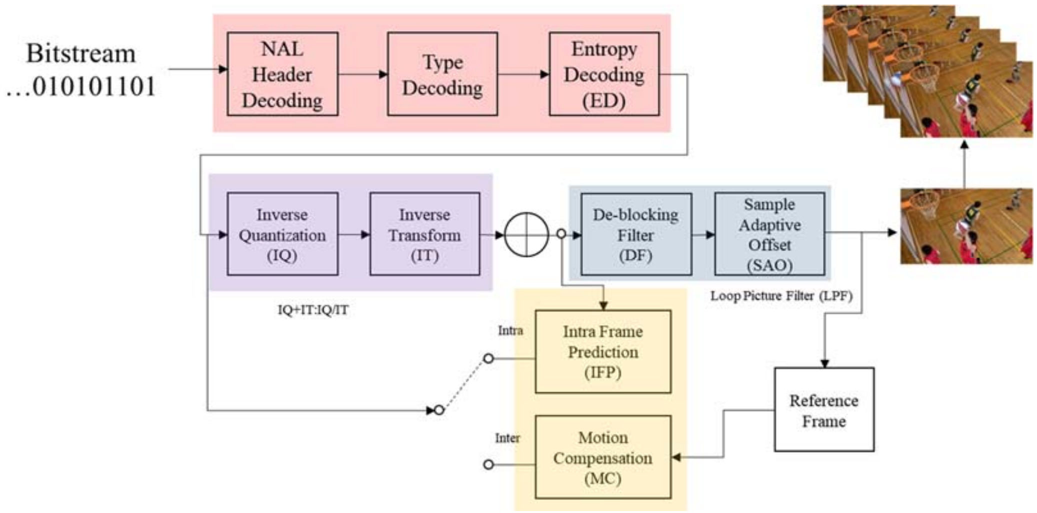


Fig. 1 HEVC decoding modules.

2.2 YOLO

YOLO is an effective real-time object detection algorithm proposed by Redmon et al [4-5]. For convenience, we introduce the first version of YOLO which is the original algorithm for object detection and location. The architecture of the YOLOv1 is shown in Fig. 2. Firstly, the input image is resized to the resolution of 448×448. And then, the well-known algorithm named convolutional neural networks (CNN) is performed to extract the feature maps of image. VGG16 [12] and GoogLeNet [13] are two outstanding CNNs which are embedded in YOLO v1, respectively. From the Fig. 2, the input image is sliced up into S × S grids [8], B bounding boxes are included in each grid cell, each with a score of confidence. The probability of an object existing in each bounding box is known as confidence according to intersection over union (IOU). The confidence (C) is defined as

$$C = Pr(object) \times IOU_{pred}^{truth} \tag{1}$$

where $Pr(object)$ is the probability of object, and IOU describes the measure of overlap between the predicted bounding box and ground truth (object), which takes the area of intersection over area of union as drawn in Fig. 3. Therefore, the IOU is defined as follows

$$IOU = \frac{area(A) \cap area(B)}{area(A) \cup area(B)} \tag{2}$$

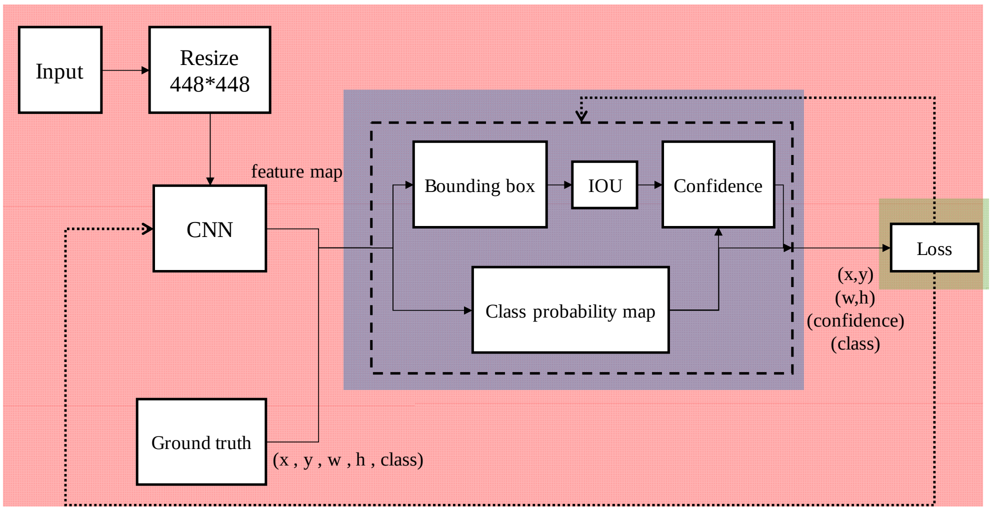


Fig. 2 The architecture of YOLO network.

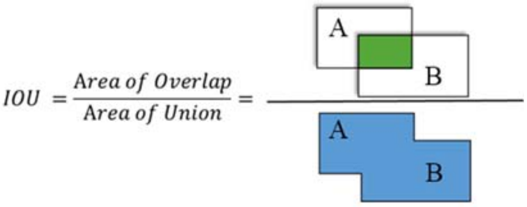


Fig. 3 Intersection over union.

When the predicted bounding box lies closer to the ground truth (IOU near to 1), YOLO can take an accurate detection. At the same time, each grid cell can predicts the conditional class probability of the object when it generates bounding boxes. Therefore, the class probability for each grid cell is derived as follows

$$Pr(Class_i|Object) \times Pr(object) \times IOU_{pred}^{truth} = Pr(Class_i) \times IOU_{pred}^{truth} \tag{3}$$

Finally, YOLO utilizes the defined loss function to train the YOLO network and improve the confidence. The loss is defined in [4] as follows

$$\begin{gathered}
\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left[(x_i-\hat{x}_i)^2 + (y_i-\hat{y}_i)^2\right] \\
+\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left[(\sqrt{w_i}-\sqrt{\widehat{w}_i})^2 + (\sqrt{h_i}-\sqrt{\widehat{h}_i})^2\right] \\
+\sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} (C_i-\hat{C}_i)^2 \\
+\lambda_{no\ obj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{no\ obj} (C_i-\hat{C}_i)^2 \\
+\textstyle\sum_{i=0}^{S^2} \mathbb{1}_{ij}^{obj} \sum_{c\in classes} (p_i(c)-\hat{p}_i(c))^2
\end{gathered} \tag{4}$$

where $\mathbb{1}_{i}^{obj}$ denotes if object appears in cell i and $\mathbb{1}_{ij}^{obj}$ denotes that the jth bounding box predictor in cell i is responsible for that prediction.

The task of YOLO object detection mainly consists of finding objects of interest in the image by marking them with bounding boxes. Bounding box is a four coordinate rectangle with a class label, which should contain the corresponding object as tightly as possible. One image can contain many possibly overlapping bounding boxes of multiple classes, such as "bike", "dog", "car", etc. [4]. In recent years, YOLO is upgraded with various versions such as YOLOv2 [5] or YOLOv3 [14] in order to optimize localization errors and increase mean average precision when applied in higher resolution videos. In this work, we will use the YOLOv2 model for object detection.

III. PROPOSED METHOD

The primary applications of YOLO series are focused on working with low-resolution images. Therefore, YOLO will lose a lot of detail in low resolution images since it is not matched original data when using high resolution capture devices. Nowadays, the IVS systems almost adopt high resolution cameras therefore bringing a need to develop a modified YOLO method to build an effective network model. In order to improve accuracy of face location for IVS of UHD videos based on HEVC, an overlapping crop method is proposed to obtain two square ratio image to overcome the limitation of image resolution.

3.1 Observation of overlapping image segmentation

Since YOLOv2 limits the resolution of input image up to square ratio of 608×608 pixels, the input UHD image needs to be scaled to a fixed size. The object detections are then resized before being fed to YOLOv2 network. As a rule of thumb, larger input images allow the detection of smaller objects but increase the computational cost and decrease the accuracy of location [12]. Figure 4 shows an example which an image with a resolution of 1280×720 is resized into 608×608. In other words, the aspect ratio of the image is changed from the original 1.77:1 (rectangular image) to 1:1 (square image). However, this will lead to the contours of the face are less obvious since the image tends to be larger and more elongated. Hence, it will occur a large reduction in the accuracy of face location for HEVC videos.

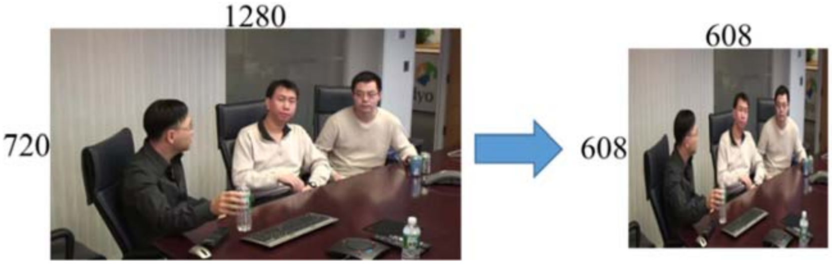


Fig. 4 An example of 1280×720 is resized into 608×608.

To solve the problem of accuracy of object detection, a direct idea is to divide rectangular image into two half images according to length. However, this method will cause some problems in the face location when faces exist in the central area. For this situation, YOLOv2 will detect two faces and take two labels, and this will increase the computational cost to determine the correct face location. In order to study a fast and precise face detection, we firstly observe the contents of each image after dividing into two images using different aspect ratios. To avoid elongation or enlargement of image after division, we proposed an overlapping segmentation method which can label the accurate face location. We divided the rectangular image into two square images according to aspect ratio, as shown in Fig. 5. Using the overlapping segmentation method, we can maintain the aspect ratios of the face in original image. Therefore, these segmented images are more suitable for the resizing method of YOLOv2.

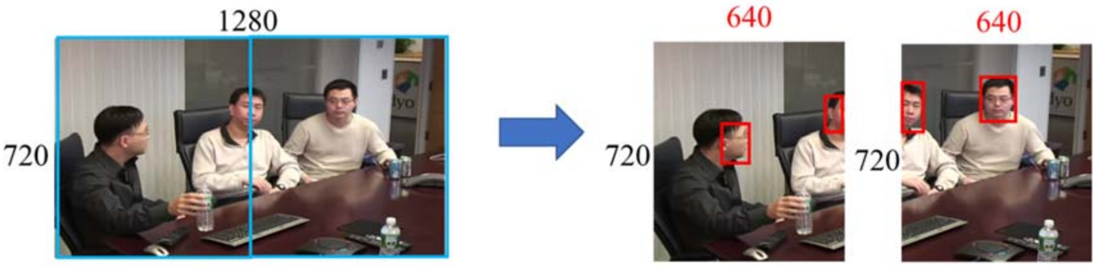
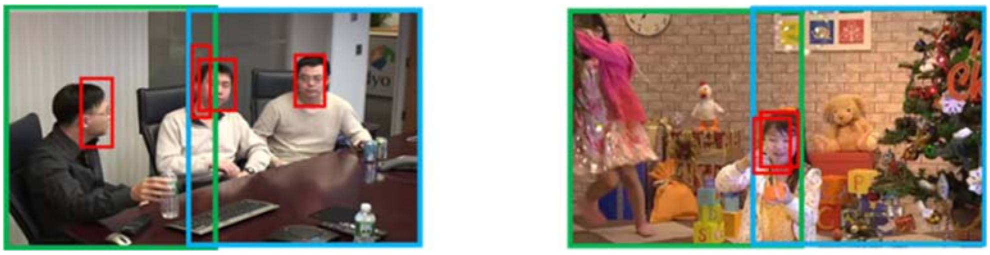


Fig. 5 Two square images according to aspect ratio.

3.2 Overlapping crop algorithm

In order to increase the accuracy of face location in UHD videos based on HEVC, we propose an overlapping crop method to obtain two square ratio image to match YOLO limitation of image resolution. Firstly, these segmented square images are separately fed to resize module, and then YOLO performs face location. However, the face in these segmented square images may be repeatedly labeled due to overlapping image segmentation. Therefore, there are two cases which will occur. In case I, two labeled boxes including one large label area with a face and one small label area with a part of face will be detected in the overlapping area, as shown in Fig. 6(a). In case II, two labeled boxes with a face will be separately detected in the overlapping area, as shown in Fig. 6(b). In the above situations, it is very difficult to delete the repeatedly labeled faces when we directly adopt the non-maximum suppression (NMS) technology [4]. This is because these values of IOU are too small to delete the repeatedly labeled face.



(a) (b)

Fig. 6 Two cases of overlapping square images. (a) Case I (b) Case II.

Therefore, in order to solve the deleting problems of repeatedly labeled faces, we will modify the IOU to match the proposed overlapping crop method. The modified IOU is written by

$$IOU' = \frac{area(A) \cap area(B)}{min(area(A), area(B))} \qquad (5)$$

where IOU' is the new accuracy measurement and $min(area(A), area(B))$ denotes the minimum area value between labeled A and labeled B. By changing the mathematical definition to calculate IOU, we can get a higher values of IOU' to match confidence and delete redundant overlapping labeled faces through NMS. The proposed overlapping crop algorithm for face location in YOLOv2 can be summarized as follows:

Step 1 Perform overlapping segmentation for UHD images from the HEVC standard dataset.
Step 2 Label faces of the segmented square image and train the YOLOv2 network.
Step 3 Two square images from segmented HEVC decoded videos are fed to YOLOv2.
Step 4 Merge two square images into one image with the same aspect ration as original HEVC image.
Step 5 Perform NMS to delete redundant overlapping labeled faces.

The procedure of the proposed overlapping crop algorithm is exhibited in Fig. 7, which takes an example of an image with resolution 1280×720 step by step.

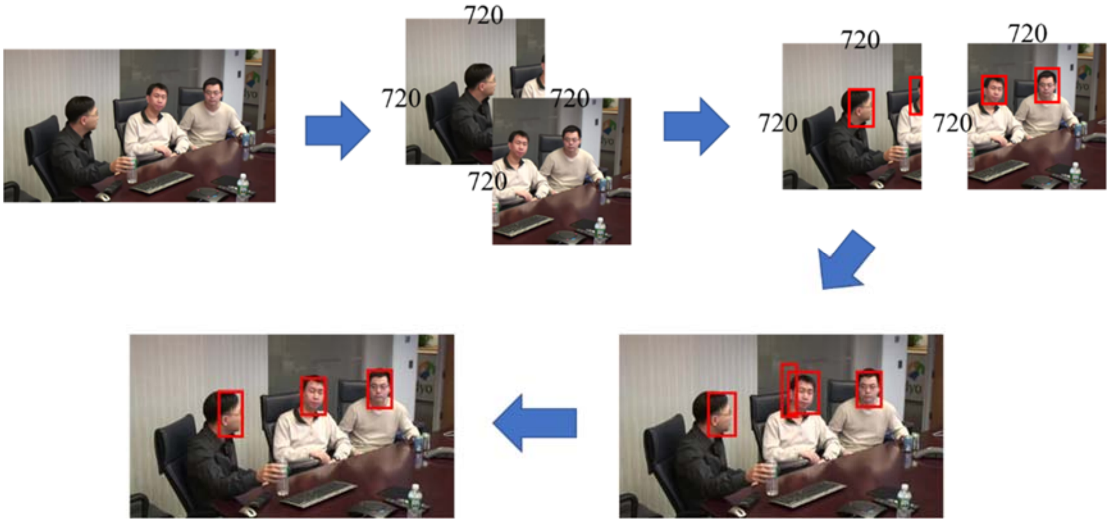


Fig. 7 The procedure of the proposed overlapping crop algorithm.

IV. EXPERIMENTAL RESULTS

In order to compare the performance of face location in UHD videos, we fed HEVC decoded images to YOLOv2 network to detect and locate a face. In our experiments, 4,800 images with different resolutions from 250×250~3840×2160 (4K), which including 3,000 training images and 1,800 testing images, were mainly taken from FDDB [15] and HEVC standard videos [16]. In addition, some real surveillance images from our laboratory are also used as training and testing dataset.

In this paper, we have decoded HEVC video bitstream in HM16.7 test model [17], the encoding configuration is All Intra with QP = 27, 32, 37. Simulations are conducted on a desktop with (1) Intel (R) Core (TW) CPU i7-3350P @ 3.60 GHz, (2) NVIDIA GeForce GTX 1060-6GB, (3) RAM: 16GB, (4) Window 10-64bit, (5) Visual Studio 2013, (5) Python 3.6.5. The accuracy performance of the proposed face location system is evaluated by using Recall measure, which is defined as follows

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FP}} \tag{6}$$

where TP is true positive and FP is false negative.

Table 1 shows the average Recall scores of face location by YOLOv2 model and the proposed system. From Table 1, we can observe that the average Recall scores of the proposed method is higher than those of original YOLOv2 model in HEVC decoded domain. Note that Recall score increases when the QP value increases, this is because decoded image quality depend on QPs. The contours of feature image become less obvious when the values of QP increase. Therefore, this leads to the accuracy drop as higher QP value. In addition, in order to evaluate the main contributions of our method for face detection and location in ultra high resolution HEVC videos. Some UHD videos including standard HEVC test videos: Class E (1280×720), Class B (1920×1080), and 4K IVS videos are implemented by the proposed method and YOLOv2 model. Table 2 shows the Recall scores of the proposed face location in UHD videos as QP=32. From Table 2, we can find when the proposed overlapping crop method is applied in 4K images, we can reach $Recall_{500} = 95\%$ and $Recall_{1000} = 95\%$ when QP=32. However, the recall rate which directly inputs 4K image to YOLOv2 model only can reach $Recall_{500} = 80\%$ and $Recall_{1000} = 83\%$, respectively. Therefore, the proposed method can finish a higher accuracy of face location in higher resolution videos.

On the other hand, to further observe the relationships between the Recall scores and image resolution, Fig. 8 shows the Recall scores vs. image resolutions. As shown in Fig. 8, the proposed method can improve the accuracy under different image resolutions. From Fig. 8, we find that the Recall score of YOLOv2 model dramatically drops as image resolution increases. However, the proposed method can keep a steady Recall score when image resolution increase from lower resolution 416×240 to ultra high resolution 3840×2160 (4K).

Table 1 Average Recall scores using different QPs.

Quantization parameter	Method	Recall_{100}	Recall_{500}	Recall_{1000}
QP=27	Proposed	94%	94%	95%
	YOLOv2	84%	85%	87%
QP=32	Proposed	92%	94%	94%
	YOLOv2	84%	86%	87%
QP=37	Proposed	92%	91%	93%
	YOLOv2	83%	83%	86%
Average	Proposed	93%	93%	94%
	YOLOv2	84%	84%	87%

Table 2 Recall scores using UHD videos.

QP=32	Method	Recall_{100}	Recall_{500}	Recall_{1000}
Class E 1280×720	Proposed	95%	95%	96%
	YOLOv2	89%	90%	90%
Class B 1920×1080	Proposed	95%	95%	95%
	YOLOv2	86%	87%	88%
4K video	Proposed	95%	95%	95%
	YOLOv2	78%	80%	83%

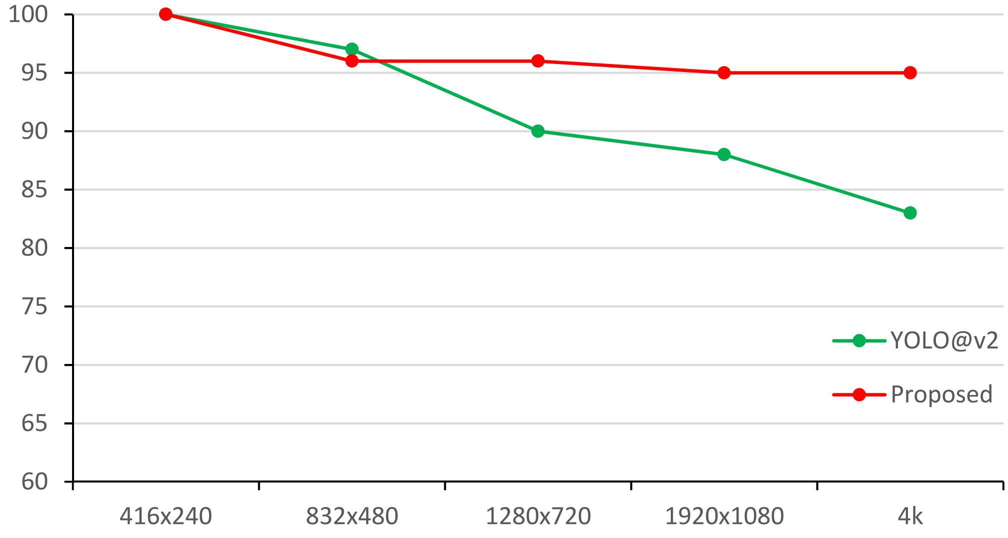


Fig. 8 The Recall scores vs. image resolutions.

V. CONCLUSION

In this paper, the proposed overlapping crop algorithm applied in face detection and location can get higher Recall scores in UHD HEVC videos as compared with those applied in YOLOv2 model. In other words, the proposed method can achieve more accurate face location. In addition, the proposed method can keep a steady Recall score when image resolution increases.

REFERENCES

[1]. G. J. Sullivan, J-R Ohm, W-J Han, T. Wiegand, “Overview of the High Efficiency Video Coding (HEVC) Standard”. IEEE Trans. on Circuitsand Systems for Video Technology, vol. 22, pp. 1649-1668, Dec. 2012.

[2]. High Efficiency Video Coding, Rec. ITU-T H.265 and ISO/IEC 23008-2, Jan. 2013.

[3]. S. J. Russel and P. Norvig, “Artificial intelligence: a modern approach,” in *Malaysia, Pearson Education Limited*, 2016.

[4]. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proc. IEEE CVPR’16*, Jun. 2016

[5]. J.Redmonand, A.Farhadi, “YOLO9000: Better, Faster, Stronger,” in *Proc. IEEE CVPR’17*, Jul. 2017.

[6]. J. Deng, W. Dong, R. Socher, L. J. Li, L. Kai and F. F. Li, “Imagenet: A large-scale hierarchical image database.” in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248- 255, 2009.

[7]. M. Everingham, L. Van Gool, C. K. Williams, J. Winn and A. Zisserman, “The PASCAL visual object classes challenge 2007 (VOC2007) results.” 2007

[8]. S. Ren, K. He, R. Girshick, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, pp. 91-99, 2015.

[9]. D. Garg, P. Goel, S. Pandya, A. Ganatra and K. Kotecha, “A Deep Learning Approach for Face Detection using YOLO,” *2018 IEEE Punecon*, 2018, pp. 1-4, doi: 10.1109/PUNECON.2018.8745376.

[10]. H. Schwarz, D. Marpe, and T. Wiegand, “Overview of the scalable video coding extension of the H.264/AVC standard,” IEEE Trans. Circuits Syst. Video Technol., vol. 17, pp. 1103-1120, Sep. 2007.

[11]. T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, “Overviewof the H.264/AVC video coding standard,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, pp. 560-576, July 2003.

[12]. K. Simonyan and A. Zisserman “Very Deep Convolutional Networks for Large-Scale Image Recognition” in *arXiv preprint arXiv:1409.1556*, 2014.

[13]. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, “Going Deeper with Convolutions” in arXiv preprint arXiv:1409.1556, 2014.

[14]. J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” arXiv preprint arXiv:1804.02767, 2018. 2

[15]. V. Jain, E. Learned-Miller, "FDDB: A Benchmark for Face Detection in Unconstrained Settings" in University of Massachusetts, 2010

[16]. YUV sequences, http://trace.eas.asu.edu/yuv/index.html

[17]. Reference software HM16.7, https://hevc.hhi. fraunhofer.de/svn/svn_HEVCSoft are/branches/